Pixal3D: Pixel-Aligned 3D Generation from Images

DONG-YANG LI, BNRist, Department of Computer Science and Technology, Tsinghua University, China
WANG ZHAO*, Tencent ARC Lab, China
YUXIN CHEN, Tencent ARC Lab, China
WENBO HU, Tencent ARC Lab, China
MENG-HAO GUO, BNRist, Department of Computer Science and Technology, Tsinghua University, China
FANG-LUE ZHANG, Victoria University of Wellington, New Zealand
YING SHAN, Tencent ARC Lab, China
SHI-MIN HU†, BNRist, Department of Computer Science and Technology, Tsinghua University, China

arXiv:2605.10922v1 [cs.CV] 11 May 2026

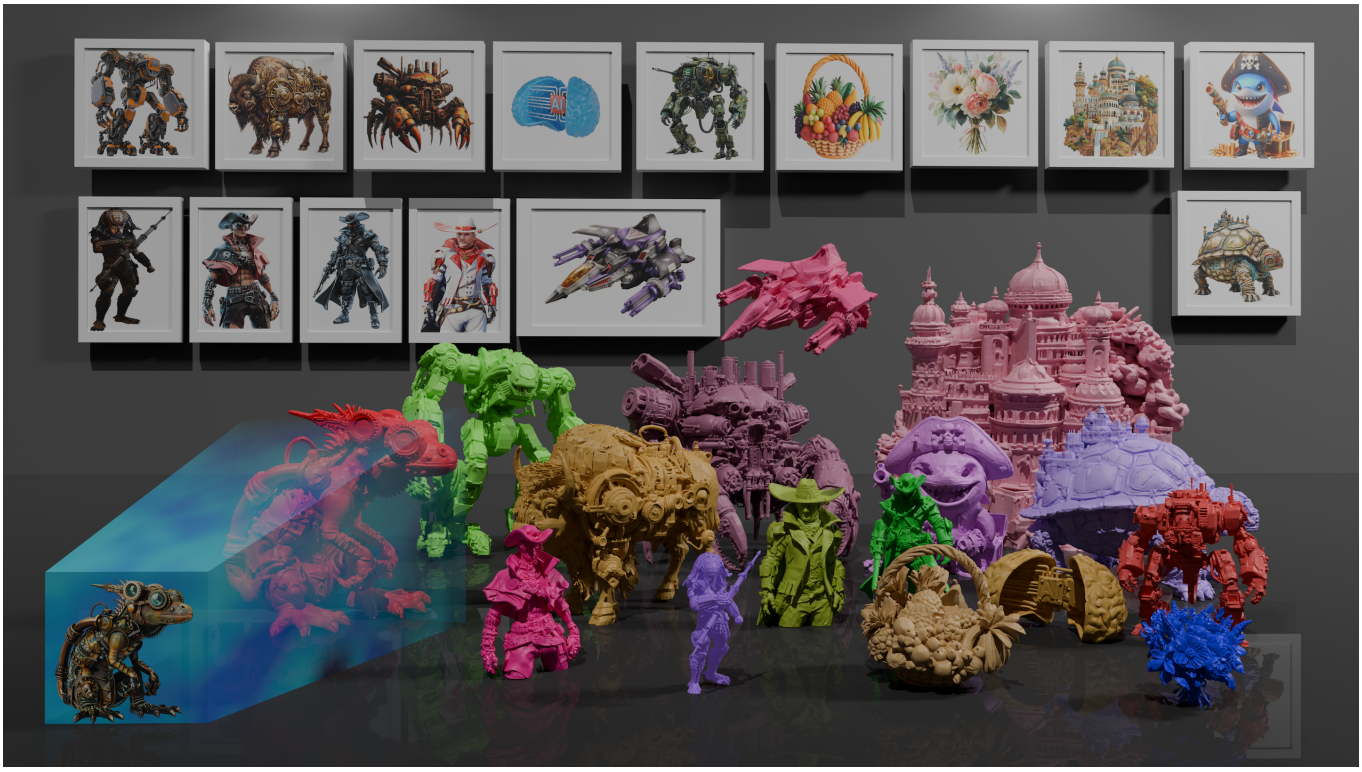

Fig. 1. Pixel-aligned meshes generated by Pixal3D. The foreground displays our results with their corresponding input images in the background. Our back-projection conditioning scheme (bottom-left) explicitly lifts 2D image features into a 3D volume to establish robust 2D-3D correspondence for generation.

*Project lead.
†Corresponding author.

Authors' Contact Information: Dong-Yang Li, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China, ldy23@mails.tsinghua.edu.cn; Wang Zhao, Tencent ARC Lab, Beijing, China, thuzhaowang@163.com; Yuxin Chen, Tencent ARC Lab, Beijing, China, chenyux53@163.com; Wenbo Hu, Tencent ARC Lab, Shenzhen, China, huwenbodut@gmail.com; Meng-Hao Guo, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China, gmh@tsinghua.edu.cn; Fang-Lue Zhang, Victoria University of Wellington, Wellington, New Zealand, z.fanglue@gmail.com; Ying Shan, Tencent ARC Lab, Shenzhen, China, yingsshan@tencent.com; Shi-Min Hu, BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing, China, shimin@tsinghua.edu.cn.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Recent advances in 3D generative models have rapidly improved image-to-3D synthesis quality, enabling higher-resolution geometry and more realistic appearance. Yet fidelity, which measures pixel-level faithfulness of the generated 3D asset to the input image, still remains a central bottleneck. We argue this stems from an implicit 2D-3D correspondence issue: most 3D-native generators synthesize shape in canonical space and inject image cues via attention, leaving pixel-to-3D associations ambiguous. To tackle this issue, we draw inspiration from 3D reconstruction and propose Pixal3D, a pixel-aligned 3D generation paradigm for high-fidelity 3D asset creation from images. Instead of generating in a canonical pose, Pixal3D directly generates

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2554-8/2026/07
https://doi.org/10.1145/3799902.3811175

3D in a pixel-aligned way, consistent with the input view. To enable this, we introduce a pixel back-projection conditioning scheme that explicitly lifts multi-scale image features into a 3D feature volume, establishing direct pixel-to-3D correspondence without ambiguity. We show that Pixal3D is not only scalable and capable of producing high-quality 3D assets, but also substantially improves fidelity, approaching the fidelity level of reconstruction. Furthermore, Pixal3D naturally extends to multi-view generation by aggregating back-projected feature volumes across views. Finally, we show pixel-aligned generation benefits scene synthesis, and present a modular pipeline that produces high-fidelity, object-separated 3D scenes from images. Pixal3D for the first time demonstrates 3D-native pixel-aligned generation at scale, and provides a new inspiring way towards high-fidelity 3D generation of object or scene from single or multi-view images. Project page: https://ldyang694.github.io/projects/pixal3d/

CCS Concepts: • **Computing methodologies** → **Mesh geometry models**; **Volumetric models**.

ACM Reference Format:
Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. 2026. Pixal3D: Pixel-Aligned 3D Generation from Images. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799902.3811175

1 Introduction

Automatic creation of high-quality 3D assets from images is a central goal in computer graphics, with profound implications for gaming, AR/VR, and digital manufacturing. Recent advances in 3D generative modeling have achieved remarkable milestones, producing assets with increasingly detailed geometry [Wu et al. 2025b; Xiang et al. 2025a], realistic appearance [Lai et al. 2025b; Yu et al. 2024] and controllable parts [Lin et al. 2025b; Yang et al. 2025b], pushing 3D generation towards truly ready-to-use assets.

However, a critical bottleneck still limits the broader adoption of current image-to-3D methods: fidelity. Here, fidelity measures how faithfully the generated 3D asset matches the input image. Most existing methods condition on an image but often produce only approximately similar shapes, with noticeable misalignment and loss of fine details. This falls short of user expectations: given an image, one typically wants the generated 3D model to (1) precisely reconstruct the visible surface, and (2) plausibly complete the unobserved regions to form a coherent and usable 3D asset. Achieving high fidelity, in addition to high quality, is a critical next step towards making image-to-3D generation genuinely useful in practice.

Interestingly, this fidelity issue is far less prominent in 3D reconstruction, a complementary field whose primary goal is to recover visible 3D structure from 2D observations, whether from multiple views or a single view. We attribute this difference to the explicit 2D-3D correspondence establishment. Correspondence is the fundamentals of reconstruction: multi-view geometry [Hartley and Zisserman 2004] is built upon pixel correspondences and triangulation, and single-view reconstruction pipelines predict depth [Lin et al. 2025a; Yang et al. 2024], normals [Fu et al. 2024; Ye et al. 2024], or point maps [Szymanowicz et al. 2025a; Wang et al. 2025c] in a pixel-aligned manner, establishing a direct, clear, one-to-one correspondence between 2D image pixels and recovered 3D. In contrast, existing 3D-native generative methods [Hunyuan3D et al. 2025; Wu et al. 2025b; Xiang et al. 2025b] synthesize shapes in a canonical pose, and rely on cross-attention to inject image information into 3D latents. This makes 2D-3D correspondence implicit and nontrivial: cross-attention must effectively "search" for where each image feature should influence the 3D representation, introducing ambiguity and confusion for local details, repetitive parts or among multiple input views, which ultimately manifests as reduced fidelity.

To resolve this fidelity issue, we propose Pixal3D, a new Pixel-Aligned 3D generation paradigm that marries the geometric rigor of reconstruction with the creative power of generative models. Unlike previous canonical space generation, Pixal3D directly generates 3D in a pixel-aligned pose consistent with the input image. To make this possible, we introduce a back-projection conditioning scheme that establishes explicit 2D-3D correspondence for injecting pixel information into 3D, replacing the commonly used cross-attention mechanism. Concretely, we back-project image features into 3D volume: every 3D voxel along that ray is assigned the corresponding pixel feature, yielding a pixel-aligned lifted 3D feature volume. This volume is then added to the 3D noise volume as a conditioning signal. We further incorporate multi-scale image features to preserve and propagate fine-grained details. Through these careful designs, we demonstrate that this pixel-aligned 3D generation paradigm is not only feasible and scalable to produce high-quality 3D models, but also significantly improves 3D fidelity over current 3D generation, achieving near reconstruction-level fidelity.

Moreover, Pixal3D naturally unifies single-view and multi-view settings under the same formulation. We extend Pixal3D to multi-view 3D generation by back-projecting each view into a pixel-aligned feature volume and aggregating them via averaging, leading to a simple and reliable multi-view generation approach. Finally, we show that this pixel-aligned paradigm also benefits 3D scene generation: we propose a modular pipeline that composes object-level generations into high-fidelity, object-separated 3D scenes, in a spirit similar to recent SAM3D [SAM et al. 2025] scene construction.

Pixal3D is essentially a 3D generative reconstruction paradigm that represents and formalizes the synergy between reconstruction and generation. It inherits the best of both worlds: the visible surfaces are tightly constrained by the input image through explicit correspondence like reconstruction, while the invisible regions are plausibly completed by learned priors of generative model conditioned on what is observed. Pixal3D provides a simple yet effective paradigm for generating faithful 3D objects and scenes from both single-view and multi-view inputs. Figure 1 shows representative examples. Importantly, Pixal3D is orthogonal to specific 3D generative backbones, and can therefore benefit from ongoing advances in geometry representations, part modeling, texturing, materials, etc., making it a scalable foundation for high-fidelity 3D generation.

Our contributions are summarized as follows: (1) We introduce Pixal3D, a pixel-aligned 3D generation paradigm, and demonstrate that pixel-aligned generation is feasible at scale while substantially improving image-to-3D fidelity. (2) We propose a ray back-projection conditioning mechanism that replaces cross-attention with explicit 2D-3D correspondence, enabling direct pixel-to-3D feature lifting and more faithful preservation of image details. (3) We extend Pixal3D from single-view to multi-view generation via simple and effective multi-view feature-volume aggregation. (4) We

propose a modular 3D scene generation pipeline based on Pixal3D that produces high-fidelity, object-separated 3D scenes.

2 Related Works

2.1 3D Generation

3D generation has advanced rapidly [Wang et al. 2025b], from distilling 2D diffusion priors into 3D [Poole et al. 2023; Wang et al. 2023] to 3D-native pipelines that learn 3D distributions from large-scale datasets [Deitke et al. 2023]. A key driver is designing 3D representations that balance fidelity, efficiency, and scalability, spanning point clouds [Nichol et al. 2022], voxels [Xiong et al. 2025], meshes [Liu et al. 2023a], 3D Gaussians [Lan et al. 2025], and triplanes [Wu et al. 2024], etc. 3DShape2VecSet [Zhang et al. 2023] introduced latent vector sets as implicit representation, later adopted and extended by [Li et al. 2025b,a,e; Zhang et al. 2024; Zhao et al. 2025] to demonstrate its scalability. To relief fidelity issue, Hi3DGen [Ye et al. 2025] introduced normal as both input and regularization. TRELLIS [Xiang et al. 2025b] proposed a sparse voxel unified representation for jointly embedding geometry and appearance, and Direct3D-S2 [Wu et al. 2025b] improved sparse voxel efficiency and regularity via spatial sparse attention. Flexible and deformable surface parameterizations are explored in Sparc3D [Li et al. 2025c] and TripoSF [He et al. 2025], enabling the generation of intricate structures and open surfaces. Inspired by Dual Contouring [Ju et al. 2002], TRELLIS 2 [Xiang et al. 2025a] and FaithC [Luo et al. 2025] incorporated dual-grid information to enhance surface representation quality. LATTICE [Lai et al. 2025a] combined compact vector sets with structural sparse voxels, proposing VoxSet for scalable generation.

Despite this progress, current state-of-the-art image-to-3D generation still faces a well-known fidelity issue: outputs are often not pixel-faithful to the input image as in reconstruction. Notably, all above methods create 3D shapes in canonical poses and condition images via cross-attention, leaving 2D-3D correspondence implicit and ambiguous, which we argue is a key cause of reduced fidelity. In contrast, Pixal3D explores a new generation paradigm to directly generate pixel-aligned 3D objects, demonstrating superior fidelity while remaining compatible with the above representation and architectural advances.

2.2 3D Reconstruction

3D reconstruction from images is a long-standing visual problem. Classical structure-from-motion (SfM) and multi-view stereo (MVS) [Schönberger and Frahm 2016; Schönberger et al. 2016] recover 3D structure by establishing correspondences, triangulation, and 2D-3D optimization such as bundle adjustment. With deep learning, approaches [Huang et al. 2018; Im et al. 2019; Yao et al. 2018] explored plane-sweeping of deep features to improve MVS robustness. Beyond 2.5D, Atlas [Murez et al. 2020] back-projects image features into a voxel grid for direct 3D prediction with 3D CNNs, and NeuralRecon [Sun et al. 2021] extends this to streaming reconstruction with similar back-projection. Our Pixal3D is inspired by these pioneers and integrates pixel-aligned back-projection into a generative backbone. Recently, feed-forward multi-view reconstruction methods like DUSt3R [Wang et al. 2024], VGGT [Wang et al. 2025a] and their followers [Tang et al. 2025; Yang et al. 2025a] have shown strong scalability by predicting pixel-aligned point maps in a shared coordinate. Similarly, single-image reconstruction has advanced including depth [Ke et al. 2024; Lin et al. 2025a; Meng et al. 2025; Yang et al. 2024; Yin et al. 2023], normal [Fu et al. 2024; Hu et al. 2024; Ye et al. 2024], point map [Wang et al. 2025c,d] or 3D Gaussian [Szymanowicz et al. 2025a,b; Zheng et al. 2024] prediction in a pixel-aligned manner.

While reconstruction recovers visible surfaces with high fidelity, its outputs are incomplete and thus not directly usable as 3D assets. Nevertheless, the explicit and unambiguous 2D-3D correspondence in reconstruction provides a key insight for generation. Pixal3D brings this principle to 3D generation via pixel-aligned modeling, enabling complete asset creation with reconstruction-level fidelity.

2.3 3D Generative Reconstruction

As 3D reconstruction and 3D generation mature, researchers increasingly realize their complementarity. This gives rise to 3D generative reconstruction, which couples reconstruction constraints with generative modeling to obtain outputs that are both consistent with inputs and complete/plausible beyond them. Early works used image generative model to complete insufficient 2D views [Liu et al. 2023b; Shi et al. 2024] to enhance reconstruction [Hong et al. 2024; Li et al. 2024]. RaySt3R [Duisterhof et al. 2025] performs ray-based novel-view prediction and fuses multi-view estimates into a complete shape, while Gen3R [Huang et al. 2025b] couples a feed-forward reconstruction backbone with diffusion to align geometry and appearance. LaRI [Li et al. 2025d] introduces view-aligned layered ray-intersection representations to better reason over occlusions. Closest to our motivation, recent works ReconViaGen [Chang et al. 2025] and CUPID [Huang et al. 2025a] target high-fidelity generative reconstruction. ReconViaGen injects VGGT features into a canonical-space generator, and CUPID jointly models a canonical 3D object and camera pose. In contrast, Pixal3D pushes this integration further and thoroughly, by establishing and enforcing explicit 2D-3D correspondence rather than predicting it: we directly generating 3d object in a pixel-aligned view-centric manner via back-projection. This design avoids the brittleness of camera estimation and reduces fidelity loss introduced by canonical-pose generation and predicted-pose dependent pixel feature fetching, leading to a scalable foundation for 3D generative reconstruction.

3 Method

Pixal3D introduces a pixel-aligned 3D generation paradigm and proposes a back-projection-based image condition scheme into a 3D latent diffusion model. This paradigm is further extended to support multi-view generation and modular scene-level synthesis. An overview of the framework is shown in Figure 2. Next, we first summarize the preliminaries of our base 3D latent diffusion model in Sec. 3.1, then detail the pixel-aligned 3D generation in Sec. 3.2, present the modular scene generation pipeline in Sec. 3.3, and discuss implementation details in Sec. 3.4.

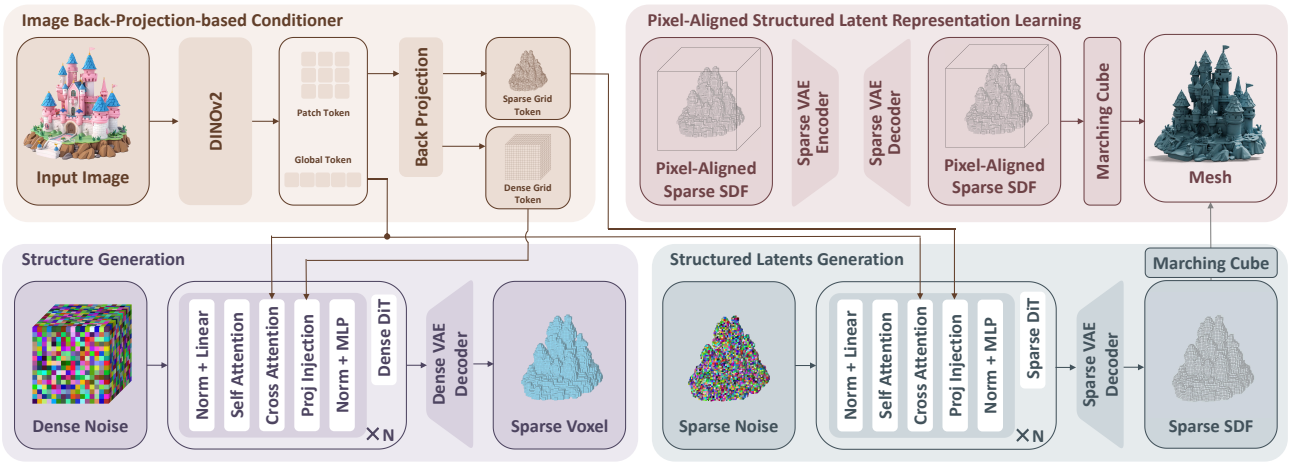

Fig. 2. Overview of the Pixal3D framework. The framework consists of three key components: (1) Pixel-Aligned Structured Latent Representation Learning (top-right), which uses a VAE to compress pixel-aligned sparse SDF into efficient sparse latents; (2) an Image Back-Projection-based Conditioner (top-left) that explicitly lifts 2D image features into 3D feature volumes; and (3) a two-stage generative process (Structure Generation and Structured Latents Generation) conditioned on these volumes to predict coarse structure and detailed latents, respectively. Finally, the generated latents are decoded into a high-fidelity mesh.

3.1 Preliminary

In principle, Pixal3D is compatible with any explicitly structured 3D generation backbone. In this work, we adopt the open-source state-of-the-art model Direct3D-S2 [Wu et al. 2025b] as our base. Direct3D-S2 is a 3D latent diffusion framework utilizing sparse voxel latents as its 3D representation. Similar to TRELLIS [Xiang et al. 2025b], it consists of a dense stage and a sparse stage, each equipped with its own VAE and DiT model. The dense stage encodes and samples a coarse occupancy grid, which is used to determine the voxel indices for the subsequent sparse stage. In the sparse stage, a sparse DiT denoises noisy sparse voxel latents, which are then decoded by a VAE decoder into a sparse SDF. Applying Marching Cubes subsequently yields the final mesh. In both the dense and sparse DiT models, image conditioning is injected via cross-attention. Pixal3D retains the core architecture of Direct3D-S2 and extends it by introducing a pixel-aligned generation paradigm.

3.2 Pixel-aligned 3D Generation

3.2.1 Canonical vs. Pixel-Aligned Generation: Existing 3D-native generation methods typically operate in an object-centric canonical pose. This representation defines a default, view-independent orientation for an object, anchoring its semantic components (e.g., a car's front, a chair's seat) to predefined axes. While this paradigm facilitates learning robust category-level priors, it fundamentally underconstrains the 2D-3D correspondence for image-conditioned generation. In practice, this correspondence is established through cross-attention between 2D and 3D tokens as a *learned behavior*. This process is inherently ambiguous: multiple 3D locations in canonical space can explain similar 2D evidence under unknown pose. Consequently, the model often cheats by using global semantic cues rather than establishing a mathematically faithful pixel-to-3D mapping.

In contrast, Pixal3D introduces *pixel-aligned generation*, where objects are defined in the input camera's coordinate frame. Intuitively, the object is represented "as seen from the camera". The generator builds view-dependent 3D behind pixels: the 3D volume is aligned with the image frustum, so each pixel corresponds to a unique camera ray and therefore a structured locus in 3D. This alignment turns correspondence from a learned, stochastic behavior into a solid geometric prior. Next, we introduce our back-projection conditioned 3D latent diffusion to realize this pixel-aligned 3D generation.

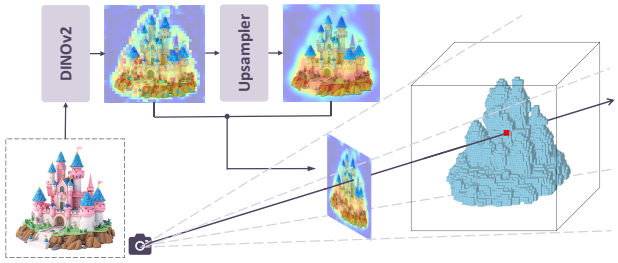

Fig. 3. Illustration of the Back-projection Conditioning Scheme.

3.2.2 Back-projection Conditioned 3D Latent Diffusion. Pixal3D is built upon 3D latent diffusion models, as introduced in Sec. 3.1. Unlike existing methods, where structured latents encoded from canonical objects serve as the diffusion target, our VAE model encodes pixel-aligned objects into 3D latents, as shown in Figure 2. Different input views thus correspond to different camera-space objects $\mathbf{X}$ and thus different latents $\mathbf{z}_0$. The diffusion model therefore learns view-dependent, pixel-aligned generation.

Back-projection condition scheme. To enable pixel-aligned generation, we introduce a back-projection scheme, instead of cross-attention, for injecting 2D image information into 3D, as shown in Figure 3. Specifically, given an input image I, we first extract a 2D feature map I' using DINOv2 [Oquab et al. 2024]. Each pixel in this feature map can be back-projected into a ray within the 3D camera coordinate system. Any 3D point along such a ray represents a potential surface point of the target object. Collectively, these rays form a camera frustum, within which the target 3D shape is assumed to reside and be defined by the image-conditioned rays.

Fig. 4. Qualitative comparisons of single view 3D generation.

The object can theoretically exist at any scale along this frustum, similar to the scale ambiguity in single-view depth estimation. However, 3D generative models often require a predefined bounding box, typically a unit cube, to specify the normalized spatial extent of the object. This cube is then voxelized (e.g., at a 64^3 resolution) to serve as input for the generative model. Therefore, we need to determine what size the cube is and where this cube should be placed within the camera frustum. We aim to ensure that the cube is not so large that the projected rays occupy only a small fraction of the voxels (degrading resolution and efficiency), yet not so small that it fails to capture the full extent of the frustum (leading to information loss). This placement is governed by a distance parameter d, which represents the distance from the camera plane to the center of the cube, and a cube scale parameter s that controls the size of the cube. With these parameters, an explicit 2D-3D correspondence can be established between image pixel (u, v) and voxel (i, j, k) inside the cube through the projection formula.

In this manner, each voxel gathers image features from its corresponding ray, forming a 3D feature volume. This feature volume provides pixel-aligned image information, which the 3D generative model uses for sampling and generation. In practice, the above process is implemented in the reverse direction following previous methods [Murez et al. 2020; Sun et al. 2021]: we project voxels onto the image plane and sample features from the image, which makes it simpler and more effective to handle interpolation. During training, we use ground-truth projection parameters, including camera intrinsics, distance d and cube scale s. For inference, we do not require these parameters; instead, we select a relatively small field of view, a unit cube scale, and then compute the camera distance such that the rays cast from the four image corners pass exactly through the four vertices of the back face of the unit cube. This ensures that the frustum information inside the cube is complete, while not sacrificing too much voxel utilization. In practice, this strategy is stable and robust, and we adopt it for all subsequent experiments.

The resulting feature volume is spatially aligned with the noise volume in the diffusion model. Therefore, we directly add the feature volume to the noise volume as the image condition. Meanwhile, we also inject the global feature token extracted by DINOv2 (image-level rather than patch-level, originally used for classification) via cross-attention, providing additional global semantic guidance.

Multi-scale 2D feature maps. While DINOv2 features contain rich image information, they are primarily composed of high-level semantic features with relatively coarse granularity. Consequently, low-level, fine-grained structural details are often lost, which to some extent constrains the fidelity of image-to-3D generation. To address this, we propose leveraging multi-scale image features to simultaneously preserve both low-level and high-level information.

Specifically, we use an off-the-shelf feature upsampling model [Chambon et al. 2025] to upscale the DINOv2 patch-token features I' to full resolution, producing a detail-rich, image-consistent map I^h. Our back-projection conditioning is unchanged: we project each voxel to the image plane, bilinearly sample features at each scale, and average them. These multi-scale high-resolution features improve the recovery and consistency of fine details. Also, it highlights the advantage of our pixel-aligned paradigm: unlike cross-attention-based methods where dense attention to high-resolution maps is prohibitively expensive, our explicit 2D-3D correspondence makes this upgrade essentially cost-free while yielding measurable gains.

3.2.3 Multi-view Extension. Since our single-view model is formulated explicitly through projection geometry, extending it to the multi-view setting is naturally straightforward. Unlike the single-view case, we assume that the camera parameters for all input views are known, consistent with the established standard setups in traditional multi-view stereo, NeRF, and 3D Gaussian Splatting, etc. Given a set of multi-view images, we back-project the multi-scale features from each view into 3D space and aggregate them within each voxel by simple averaging. The resulting fused feature volume is then used as the conditioning signal for the generative model. This simple yet effective multi-view conditioning scheme accommodates an arbitrary number of input views. As the number of viewpoints increases, a greater extent of the 3D surface becomes visible, leading to a more deterministic 3D shape reconstruction.

Fig. 5. Qualitative comparison of single-view 3D generation on in-the-wild images.

Fig. 6. Qualitative comparison of multi-view 3D generation on Toys4K.

Table 1. Quantitative evaluation of single-view generation on Toys4K. Metrics compare rendered and ground-truth normals.

Method	IoU↑	PSNR↑	SSIM↑	LPIPS↓	Mean↓	Median↓	Mean_B↓	11.25°↑	22.5°↑	30°↑
TRELLIS [Xiang et al. 2025b]	79.48	20.98	0.883	0.204	25.00	17.97	36.04	46.82	63.99	70.80
TripoSG [Li et al. 2025e]	73.54	19.73	0.873	0.250	28.55	21.20	41.71	39.85	57.18	64.81
Hunyuan3D-2.1 [Hunyuan3D et al. 2025]	83.33	21.96	0.889	0.179	21.19	14.05	32.46	51.37	69.08	75.83
Direct3D-S2 [Wu et al. 2025b]	74.23	19.49	0.851	0.268	29.99	23.46	41.04	37.56	55.46	63.20
Pixal3D	**93.57**	**24.21**	**0.897**	**0.108**	**16.63**	**11.77**	**21.80**	**53.13**	**77.96**	**85.35**

3.3 Scene Generation Pipeline

Pixal3D enables pixel-aligned 3D object generation from images. Consequently, when the input is a scene image containing multiple objects, Pixal3D can generate each object individually and then compose them via image-space alignment, yielding a full 3D scene synthesis. A recent representative work for this task is SAM3D [SAM et al. 2025]. SAM3D first leverages SAM3 [Carion et al. 2025] to interactively segment objects in image, and then trains an object generator on TRELLIS backbone. Given an RGB crop of a partially visible, occluded object, it generates the object in a canonical pose and predicts its pose (rotation, translation, and scale) in the camera frame to align objects into a consistent, object-separated 3D scene.

In contrast, Pixal3D generates each object directly in the camera space in a pixel-aligned manner, which makes multi-object alignment substantially simpler. We propose a modular scene generation pipeline comprising three steps: (1) *Segmentation and Completion:* we employ SAM3 [Carion et al. 2025] for interactive segmentation to obtain object masks, followed by Qwen-image-edit [Wu et al. 2025a] to perform 2D completion of occluded regions. (2) *Pixel-Aligned Generation:* These completed images are fed into Pixal3D for 3D generation. Since the orientation of our generated objects is already aligned with the input image, we only need to resolve relative scale and depth across objects. (3) *Global Alignment:* we use MoGe [Wang et al. 2025c] to predict a global point map from image. The pixel-aligned nature of both Pixal3D's outputs and MoGe's predictions allow us to directly formulate point-wise constraints and solve a least-squares problem to estimate object scale and depth.

Compared to SAM3D, this pipeline offers superior fidelity and geometric detail in single-object generation. Furthermore, our pipeline avoids the challenging and often non-robust step of estimating a 7-DoF object pose from the image. Instead, we resolve alignment through pixel-aligned generation and global depth estimation, which substantially improves both the accuracy and stability of multi-object alignment. Our pipeline yields higher-fidelity 3D scene generation results, providing a promising alternative perspective for holistic 3D scene generation from a single image.

3.4 Implementation Details

To train Pixal3D, we use the TRELLIS-500K [Xiang et al. 2025b] subset of Objaverse dataset [Deitke et al. 2023]. To construct pixel-aligned image-mesh pairs, we apply random object-centric rotations and render meshes from frontal perspectives with varying FoVs and camera distances. We watertight each mesh and compute its SDF. For model architecture, we use the same VAE and DiT model as Direct3D-S2 [Wu et al. 2025b], except that we replace the cross-attention conditioning with our back-projection-based conditioner.

Table 2. Quantitative evaluation of single-view generation on in-the-wild test set. User study collects scores for both fidelity and quality.

Method	Uni3D↑	ULIP2↑	Fidelity↑	Quality↑
TRELLIS	41.09	44.76	1.86	1.99
TripoSG	40.99	44.64	2.25	2.14
Hunyuan3D-2.1	41.15	44.65	2.77	2.50
Direct3D-S2	41.62	44.79	3.21	3.64
Pixal3D	**42.11**	**45.04**	**4.91**	**4.74**

The pretrained VAE works robustly for pixel-aligned SDF, thus we only finetune the decoder for better quality. For sparse DiT, following Direct3D-S2, we adopt a coarse-to-fine training schedule, training at resolutions of 256, 384, 512, and 1024 for 200k, 100k, 80k, and 40k iterations, respectively. The dense DiT is trained for 300k iterations with a learning rate of 1e-4, followed by an additional 200k iterations at 2e-5. For image conditioning, we use DINOv2-Large as encoder, and employ NAF [Chambon et al. 2025] to upsample to the DINOv2's input resolution of 518x518. For multi-view training, the pre-trained single-view model is fine-tuned, with random sampling of 2 to 6 views as condition. Code will be released publicly.

4 Experiments

4.1 Single-view 3D Generation

To validate the effectiveness of our Pixal3D framework, we conduct comprehensive quantitative and qualitative evaluations against representative state-of-the-art 3D generation methods, including TRELLIS [Xiang et al. 2025b], TripoSG [Li et al. 2025e], Hunyuan3D-2.1 [Hunyuan3D et al. 2025], and Direct3D-S2 [Wu et al. 2025b].

Quantitative Comparison. To precisely assess fidelity differences across methods, we render the surface normals of each generated 3D mesh in the input image coordinate frame, then compare these with ground-truth normal maps. This evaluation is performed on all meshes in the Toys4K dataset [Stojanov et al. 2021]. For the baselines, we use the ground-truth camera pose for normal rendering. In contrast, owing to its pixel-aligned nature, our method directly leverages its inference-time projection for normal rendering.

As for metrics, we use IoU to measure the overlap between the rendered and the ground-truth normal maps, and PSNR to quantify their pixel-wise discrepancy. In addition, we report commonly used error metrics in normal estimation: mean and median angular error, mean angular error around image boundaries (Mean_B), and accuracy under different angular thresholds. All these metrics are computed only on the overlapping regions where both prediction

Table 3. Quantitative evaluation of multi-view generation on Toys4K.

Method	View = 2			View = 4			View = 6		
	CD (10^{-4})↓	EMD (10^{-2})↓	F-Score↑	CD (10^{-4})↓	EMD (10^{-2})↓	F-Score↑	CD (10^{-4})↓	EMD (10^{-2})↓	F-Score↑
VGGT	613.55	19.60	9.57	881.53	21.71	10.25	2791.10	25.33	9.67
TRELLIS	21.39	2.40	43.68	18.35	2.19	45.90	18.13	2.16	46.02
Pixal3D	**5.27**	**1.13**	**64.94**	**4.73**	**1.05**	**67.85**	**4.16**	**1.00**	**69.04**

and ground truth are available. The results are summarized in Table 1. Our method achieves substantial improvements across all metrics, demonstrating significantly better fidelity.

Since the meshes in Toys4K are mostly simple, we further collect 150 images from the Internet and AI-generated sources as an additional test set, featuring complex geometric details and diverse semantics. On this test set, given the absence of ground-truth camera poses or normal maps, we evaluate image-3D consistency using ULIP2 [Xue et al. 2024] and Uni3D [Zhou et al. 2024]. We also conduct a user study with 30 participants on this test set. Participants are asked to score the generated meshes from two aspects: fidelity (image-3D consistency) and quality (overall 3D shape quality) from 1 (worst) to 5 (best). These results are summarized in Table 2. Our method is favored by the majority of participants, especially in terms of fidelity, highlighting its superior ability to faithfully preserve image details while maintaining high overall quality.

Qualitative comparisons. Figure 4 and 5 presents visual comparison examples. Compared to all other methods, our approach more faithfully and accurately recovers the visual content of input image and produces higher-quality 3D meshes. The fidelity gap is particularly evident in fine-grained details, such as keyboard layouts, facial details including eyes, the number and arrangement of flower petals, etc. These examples illustrate the misalignment caused by 2D-3D correspondence ambiguity in prior methods. In contrast, thanks to its pixel-aligned formulation, our method preserves nearly all image details, achieving an almost reconstruction-level fidelity. More examples are provided in the supplementary material and the video.

4.2 Multi-view 3D Generation

For multiview evaluation, we select representative baselines from both multiview reconstruction and generation, specifically VGGT [Wang et al. 2025a] and TRELLIS (multiview version) [Xiang et al. 2025b]. Evaluations are conducted on the Toys4k dataset using Chamfer Distance (CD), Earth Mover's Distance (EMD), and F-Score. We evaluate performance with varying numbers of input views (2, 4, and 6). Table 3 summarizes the results. Our method significantly outperforms the baseline approaches across all metrics.

Qualitative results are presented in Figure 6. Specifically, VGGT often fails to produce strictly aligned point cloud reconstructions and frequently exhibits significant floaters and outliers. While the multi-view variant of TRELLIS produces smooth mesh outputs, its multi-view fidelity remains limited. It struggles to ensure consistency across all views and occasionally introduces hallucinations. In contrast, our pixel-aligned formulation seamlessly accommodates multi-view inputs, resulting in superior cross-view consistency. Moreover, as the number of views increases, generative ambiguity decreases while reconstruction cues become stronger, a trend consistently observed in our results. This behavior is also a fundamental principle and objective of 3D generative reconstruction.

Fig. 7. Qualitative comparison on 3D scene generation.

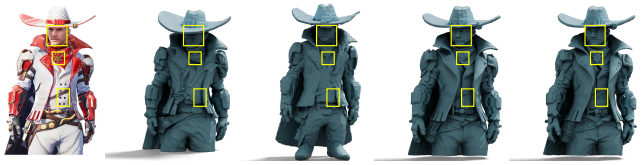

Fig. 8. Ablation study on key components.

4.3 3D Scene Generation

We extend Pixal3D to scene generation, as discussed in Sec. 3.3. Figure 7 presents a qualitative comparison between our results and recent SAM3D [SAM et al. 2025]. SAM3D jointly estimates canonical geometry and object poses, but its per-object pose estimation often yields inconsistent, non-robust inter-object relations (e.g., wrong relative rotations, misaligned placements, and incorrect contact/support), as shown in the figure. In contrast, our method enforces pixel-aligned constraints for each object relative to the input image and regularizes their spatial consistency through geometric cues like depth maps, leading to more coherent and practically usable scene generation results.

4.4 Ablation Studies

We conduct ablation studies to validate the effectiveness of our key modules. The results are shown in Figure 8. Without feature upsampling, 3D generation must rely on relatively coarse feature maps (e.g., 37×37 patch tokens from DINOv2) to represent image details. This inevitably leads to missing fine details and misalignment in the generated 3D results, as illustrated in the figure. Furthermore, when the back-projection conditioning scheme is removed and replaced by a conventional cross-attention mechanism for a pixel-aligned 3D generator, training becomes slow to converge and unstable, and the

final results exhibit substantially lower fidelity. These observations highlight the necessity of our design choices.

4.5 Limitations and Future Works

While our method demonstrates strong 3D generative reconstruction performance, several limitations remain. First, the framework exhibits sensitivity to pixel-level noise (e.g. imperfect segmentation boundaries), which can be back-projected and amplified into small geometric artifacts. Second, our current multi-view formulation assumes known and reasonably accurate camera poses. Third, our scene-generation pipeline relies on 2D inpainting to complete occluded regions, which may occasionally introduce errors in complex occlusions. Moving forward, a natural next step is to extend the current geometry backbone with texture and material synthesis, where our pixel-aligned paradigm is particularly suited to improve appearance fidelity. In addition, pixel-aligned generation opens opportunities for downstream 3D editing and interaction via 2D pixel manipulation. Finally, extending pixel-aligned generation to video-based 3D scene generation would be an interesting direction, bridging high-fidelity asset creation with controllable world building.

5 Conclusion

In this paper, we present Pixal3D, a pixel-aligned 3D generation paradigm for high-fidelity 3D asset creation from images. Unlike existing 3D-native generation methods that synthesize shapes in canonical space, Pixal3D directly creates 3D models aligned with images. A back-projection based image condition scheme replaces ambiguous cross-attention with explicit, geometric 2D-3D correspondence, enabling high-precision, pixel-aligned synthesis. We further demonstrated the versatility of this paradigm by extending it to multi-view inputs and 3D scene generation through a modular pipeline. Our extensive evaluations confirm that pixel-aligned generation is not only feasible but significantly enhances 3D fidelity. Pixal3D provides a scalable foundation for 3D generative reconstruction, offering a promising path towards creating 3D content that is both creatively flexible and pixel-faithful.

Acknowledgments

This work was supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101), the National Natural Science Foundation of China (project No. 62495060), the Research Grant of Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. 2025. SAM 3: Segment Anything with Concepts. *CoRR* abs/2511.16719 (2025). arXiv:2511.16719 doi:10.48550/ARXIV.2511.16719

Loïck Chambon, Paul Couairon, Eloi Zablocki, Alexandre Boulch, Nicolas Thome, and Matthieu Cord. 2025. NAF: Zero-Shot Feature Upsampling via Neighborhood Attention Filtering. *CoRR* abs/2511.18452 (2025). arXiv:2511.18452 doi:10.48550/ARXIV.2511.18452

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. 2025. ReconViaGen: Towards Accurate Multi-view 3D Object Reconstruction via Generation. *CoRR* abs/2510.23306 (2025). arXiv:2510.23306 doi:10.48550/ARXIV.2510.23306

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/70364304877b5e767de4e9a2a511be0c-Abstract-Datasets_and_Benchmarks.html

Bardienus Pieter Duisterhof, Jan Oberst, Bowen Wen, Stan Birchfield, Deva Ramanan, and Jeffrey Ichnowski. 2025. RaySt3R: Predicting Novel Depth Maps for Zero-Shot Object Completion. *CoRR* abs/2506.05285 (2025). arXiv:2506.05285 doi:10.48550/ARXIV.2506.05285

Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. 2024. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXII (Lecture Notes in Computer Science, Vol. 15080)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 241–258. doi:10.1007/978-3-031-72670-5_14

Richard Hartley and Andrew Zisserman. 2004. *Multiple View Geometry in Computer Vision.* Cambridge University Press. doi:10.1017/CBO9780511811685

Xianglong He, Zi-Xin Zou, Chia-Hao Chen, Yuan-Chen Guo, Ding Liang, Chun Yuan, Wanli Ouyang, Yan-Pei Cao, and Yangguang Li. 2025. SparseFlex: High-Resolution and Arbitrary-Topology 3D Shape Modeling. *CoRR* abs/2503.21732 (2025). arXiv:2503.21732 doi:10.48550/ARXIV.2503.21732

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. 2024. LRM: Large Reconstruction Model for Single Image to 3D. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=sllU8vvsFF

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 2024. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 46, 12 (2024), 10579–10596. doi:10.1109/TPAMI.2024.3444912

Binbin Huang, Haobin Duan, Yiqun Zhao, Zibo Zhao, Yi Ma, and Shenghua Gao. 2025a. CUPID: Pose-Grounded Generative 3D Reconstruction from a Single Image. *CoRR* abs/2510.20776 (2025). arXiv:2510.20776 doi:10.48550/ARXIV.2510.20776

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. 2025b. Gen3R: 3D Scene Generation Meets Feed-Forward Reconstruction. *CoRR* abs/2601.04090 (2025). arXiv:2601.04090 doi:10.48550/ARXIV.2601.04090

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. 2018. DeepMVS: Learning Multi-View Stereopsis. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2821–2830. doi:10.1109/CVPR.2018.00298

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, Qingxiang Lin, Zeqiang Lai, Xianghui Yang, Huiwen Shi, Zibo Zhao, Bowen Zhang, Hongyu Yan, Lifu Wang, Sicong Liu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Dongyuan Guo, Junlin Yu, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Shida Wei, Chao Zhang, Yonghao Tan, Yifu Sun, Lin Niu, Shirui Huang, Bojian Zheng, Shu Liu, Shilin Chen, Xiang Yuan, Xiaofeng Yang, Kai Liu, Jianchen Zhu, Peng Chen, Tian Liu, Di Wang, Yuhong Liu, Linus, Jie Jiang, Jingwei Huang, and Chunchao Guo. 2025. Hunyuan3D 2.1: From Images to High-Fidelity 3D Assets with Production-Ready PBR Material. *CoRR* abs/2506.15442 (2025). arXiv:2506.15442 doi:10.48550/ARXIV.2506.15442

Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. 2019. DPSNet: End-to-end Deep Plane Sweep Stereo. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=ryeYHi0ctQ

Tao Ju, Frank Losasso, Scott Schaefer, and Joe D. Warren. 2002. Dual contouring of hermite data. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 339–346. doi:10.1145/566654.566586

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2024. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 9492–9502. doi:10.1109/CVPR52733.2024.00907

Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Haolin Liu, Qingxiang Lin, Jingwei Huang, Chunchao Guo, and Xiangyu Yue. 2025a. Faithful Contouring: Near-Lossless 3D Voxel

Representation Free from Iso-surface. *CoRR* abs/2512.03052 (2025). arXiv:2512.03052 doi:10.48550/ARXIV.2512.03052

Zeqiang Lai, Yunfei Zhao, Zibo Zhao, Xin Yang, Xin Huang, Jingwei Huang, Xiangyu Yue, and Chunchao Guo. 2025b. NaTex: Seamless Texture Generation as Latent Color Diffusion. *CoRR* abs/2511.16317 (2025). arXiv:2511.16317 doi:10.48550/ARXIV.2511.16317

Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. 2025. GaussianAnything: Interactive Point Cloud Flow Matching for 3D Generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net. https://openreview.net/forum?id=P4DbTSDQFu

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. 2024. Instant3D: Fast Text-to-3D with Sparse-view Generation and Large Reconstruction Model. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=2lDQLiH1W4

Rui Li, Biao Zhang, Zhenyu Li, Federico Tombari, and Peter Wonka. 2025d. LaRI: Layered Ray Intersections for Single-view 3D Geometric Reasoning. *CoRR* abs/2504.18424 (2025). arXiv:2504.18424 doi:10.48550/ARXIV.2504.18424

Weiyu Li, Jiarui Liu, Hongyu Yan, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. 2025b. CraftsMan3D: High-fidelity Mesh Generation with 3D Native Diffusion and Interactive Geometry Refiner. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 5307–5317. doi:10.1109/CVPR52734.2025.00500

Xiao-Lei Li, Hao-Xiang Chen, Yanni Zhang, Kai Ma, Alan Zhao, Tai-Jiang Mu, Hao-Xiang Guo, and Ran Zhang. 2025a. RELATE3D: REfocusing Latent Adapter for Targeted local Enhancement and Editing in 3D Generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference, SIGGRAPH Conference Papers 2025, Vancouver, BC, Canada, August 10-14, 2025*, Ginger Alford, Hao (Richard) Zhang, and Adriana Schulz (Eds.). ACM, 79:1–79:12. doi:10.1145/3721238.3730648

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, and Yan-Pei Cao. 2025e. TripoSG: High-Fidelity 3D Shape Synthesis using Large-Scale Rectified Flow Models. *CoRR* abs/2502.06608 (2025). arXiv:2502.06608 doi:10.48550/ARXIV.2502.06608

Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. 2025c. Sparc3D: Sparse Representation and Construction for High-Resolution 3D Shapes Modeling. *CoRR* abs/2505.14521 (2025). arXiv:2505.14521 doi:10.48550/ARXIV.2505.14521

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. 2025a. Depth Anything 3: Recovering the Visual Space from Any Views. *CoRR* abs/2511.10647 (2025). arXiv:2511.10647 doi:10.48550/ARXIV.2511.10647

Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. 2025b. PartCrafter: Structured 3D Mesh Generation via Compositional Latent Diffusion Transformers. *CoRR* abs/2506.05573 (2025). arXiv:2506.05573 doi:10.48550/ARXIV.2506.05573

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023b. Zero-1-to-3: Zero-shot One Image to 3D Object. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 9264–9275. doi:10.1109/ICCV51070.2023.00853

Zhen Liu, Yao Feng, Michael J. Black, Derek Nowrouzezahrai, Liam Paull, and Weiyang Liu. 2023a. MeshDiffusion: Score-based Generative 3D Mesh Modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=0cpM2ApF9p6

Yihao Luo, Xianglong He, Chuanyu Pan, Yiwen Chen, Jiaqi Wu, Yangguang Li, Wanli Ouyang, Yuanming Hu, Guang Yang, and Choon Hwai Yap. 2025. Faithful Contouring: Near-Lossless 3D Voxel Representation Free from Iso-surface. *CoRR* abs/2511.04029 (2025). arXiv:2511.04029 doi:10.48550/ARXIV.2511.04029

Ming Meng, Yonggui Zhu, Yufei Zhao, Zhaoxin Li, and Zhe Zhu. 2025. 3D Indoor Scene Geometry Estimation from a Single Omnidirectional Image: A Comprehensive Survey. *Computational Visual Media* 11, 3 (2025), 431–464. doi:10.26599/CVM.2025.9450438

Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. 2020. Atlas: End-to-End 3D Scene Reconstruction from Posed Images. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 12352)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 414–431. doi:10.1007/978-3-030-58571-6_25

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *CoRR* abs/2212.08751 (2022). arXiv:2212.08751 doi:10.48550/ARXIV.2212.08751

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.* 2024 (2024). https://openreview.net/forum?id=a68SUt6zFt

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2023. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=FjNys5c7VyY

SAM, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. 2025. SAM 3D: 3Dfy Anything in Images. *CoRR* abs/2511.16624 (2025). arXiv:2511.16624 doi:10.48550/ARXIV.2511.16624

Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 4104–4113. doi:10.1109/CVPR.2016.445

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 9907)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 501–518. doi:10.1007/978-3-319-46487-9_31

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. 2024. MVDream: Multi-view Diffusion for 3D Generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=FUgrjq2pbB

Stefan Stojanov, Anh Thai, and James M. Rehg. 2021. Using Shape To Categorize: Low-Shot Learning With an Explicit Shape Bias. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 1798–1808. doi:10.1109/CVPR46437.2021.00184

Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. 2021. NeuralRecon: Real-Time Coherent 3D Reconstruction From Monocular Video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 15598–15607. doi:10.1109/CVPR46437.2021.01534

Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F. Henriques, Christian Rupprecht, and Andrea Vedaldi. 2025a. Flash3D: Feed-Forward Generalisable 3D Scene Reconstruction from a Single Image. In *International Conference on 3D Vision, 3DV 2025, Singapore, March 25-28, 2025*. IEEE, 670–681. doi:10.1109/3DV66043.2025.00067

Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. 2025b. Bolt3d: Generating 3d scenes in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 24846–24857.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander G. Schwing, and Zhicheng Yan. 2025. MV-DUSt3R+: Single-Stage Scene Reconstruction from Sparse Views In 2 Seconds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 5283–5293. doi:10.1109/CVPR52734.2025.00498

Chen Wang, Hao-Yang Peng, Ying-Tian Liu, Jiatao Gu, and Shi-Min Hu. 2025b. Diffusion Models for 3D Generation: A Survey. *Computational Visual Media* 11, 1 (2025), 1–28. doi:10.26599/CVM.2025.9450452

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. 2025a. VGGT: Visual Geometry Grounded Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 5294–5306. doi:10.1109/CVPR52734.2025.00499

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2025c. MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 5261–5271. doi:10.1109/CVPR52734.2025.00496

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. 2025d. MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details. *CoRR* abs/2507.02546 (2025). arXiv:2507.02546 doi:10.48550/ARXIV.2507.02546

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 20697–20709. doi:10.1109/CVPR52733.2024.01956

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/1a87980b9853e84dfb295855b425c262-Abstract-Conference.html

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025a. Qwen-Image Technical Report. *CoRR* abs/2508.02324 (2025). arXiv:2508.02324 doi:10.48550/ARXIV.2508.02324

Shuang Wu, Youtian Lin, Yifei Zeng, Feihu Zhang, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. 2024. Direct3D: Scalable Image-to-3D Generation via 3D Latent Diffusion Transformer. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/dc970c91c0a82c6e4cb3c4af7bff5388-Abstract-Conference.html

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, and Yao Yao. 2025b. Direct3D-S2: Gigascale 3D Generation Made Easy with Spatial Sparse Attention. *CoRR* abs/2505.17412 (2025). arXiv:2505.17412 doi:10.48550/ARXIV.2505.17412

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. 2025a. Native and Compact Structured Latents for 3D Generation. *CoRR* abs/2512.14692 (2025). arXiv:2512.14692 doi:10.48550/ARXIV.2512.14692

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. 2025b. Structured 3D Latents for Scalable and Versatile 3D Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 21469–21480. doi:10.1109/CVPR52734.2025.02000

Bojun Xiong, Si-Tong Wei, Xin-Yang Zheng, Yan-Pei Cao, Zhouhui Lian, and Peng-Shuai Wang. 2025. OctFusion: Octree-based Diffusion Models for 3D Shape Generation. *Comput. Graph. Forum* 44, 5 (2025). doi:10.1111/CGF.70198

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. 2024. ULIP-2: Towards Scalable Multimodal Pre-Training for 3D Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 27081–27091. doi:10.1109/CVPR52733.2024.02558

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. 2025a. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 21924–21935. doi:10.1109/CVPR52734.2025.02042

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 10371–10381. doi:10.1109/CVPR52733.2024.00987

Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. 2025b. OmniPart: Part-Aware 3D Generation with Semantic Decoupling and Structural Cohesion. *CoRR* abs/2507.06165 (2025). arXiv:2507.06165 doi:10.48550/ARXIV.2507.06165

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. 2018. MVSNet: Depth Inference for Unstructured Multi-view Stereo. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII (Lecture Notes in Computer Science, Vol. 11212)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer, 785–801. doi:10.1007/978-3-030-01237-3_47

Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. 2024. StableNormal: Reducing Diffusion Variance for Stable and Sharp Normal. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 250:1–250:18. doi:10.1145/3687971

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. 2025. Hi3DGen: High-fidelity 3D Geometry Generation from Images via Normal Bridging. *CoRR* abs/2503.22236 (2025). arXiv:2503.22236 doi:10.48550/ARXIV.2503.22236

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 9009–9019. doi:10.1109/ICCV51070.2023.00830

Xin Yu, Ze Yuan, Yuan-Chen Guo, Ying-Tian Liu, Jianhui Liu, Yangguang Li, Yan-Pei Cao, Ding Liang, and Xiaojuan Qi. 2024. TEXGen: a Generative Diffusion Model for Mesh Textures. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 213:1–213:14. doi:10.1145/3687909

Biao Zhang, Jiapeng Tang, Matthias Nießner, and Peter Wonka. 2023. 3DShape2VecSet: A 3D Shape Representation for Neural Fields and Generative Diffusion Models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 92:1–92:16. doi:10.1145/3592442

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. 2024. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 120:1–120:20. doi:10.1145/3658146

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, Huiwen Shi, Sicong Liu, Junta Wu, Yihang Lian, Fan Yang, Ruining Tang, Zebin He, Xinzhou Wang, Jian Liu, Xuhui Zuo, Zhuo Chen, Biwen Lei, Haohan Weng, Jing Xu, Yiling Zhu, Xinhai Liu, Lixin Xu, Changrong Hu, Tianyu Huang, Lifu Wang, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Yulin Cai, Jiaao Yu, Yixuan Tang, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Chao Zhang, Yonghao Tan, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Zhichao Hu, Lei Qin, Jianbing Peng, Zhan Li, Minghui Chen, Xipeng Zhang, Lin Niu, Paige Wang, Yingkai Wang, Haozhao Kuang, Zhongyi Fan, Xu Zheng, Weihao Zhuang, YingPing He, Tian Liu, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, Jingwei Huang, and Chunchao Guo. 2025. Hunyuan3D 2.0: Scaling Diffusion Models for High Resolution Textured 3D Assets Generation. *CoRR* abs/2501.12202 (2025). arXiv:2501.12202 doi:10.48550/ARXIV.2501.12202

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2024. GPS-Gaussian: Generalizable Pixel-Wise 3D Gaussian Splatting for Real-Time Human Novel View Synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 19680–19690. doi:10.1109/CVPR52733.2024.01861

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. 2024. Uni3D: Exploring Unified 3D Representation at Scale. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=wcaE4Dfgt8